

A Real-time Continuous Gesture Recognition System for Sign Language

Rung-Huei Liang¹, Ming Ouhyoung²

¹Dept. of Information Management, Shih-Chien University, Taichih, Taipei 104, Taiwan, R.O.C.
liang@scc1.scc.edu.tw

²Dept. of Computer Science and Information Engineering, National Taiwan University, Taipei 106
Taiwan, R.O.C.; ming@csie.ntu.edu.tw

Abstract

In this paper, a large vocabulary sign language interpreter is presented with real-time continuous gesture recognition of sign language using a DataGloveTM. The most critical problem, end-point detection in a stream of gesture input is first solved and then statistical analysis is done according to 4 parameters in a gesture : posture, position, orientation, and motion. We have implemented a prototype system with a lexicon of 250 vocabularies in Taiwanese Sign Language (TWL). This system uses hidden Markov models (HMMs) for 51 fundamental postures, 6 orientations, and 8 motion primitives. In a signer-dependent way, a sentence of gestures based on these vocabularies can be continuously recognized in real-time and the average recognition rate is 80.4%.

1. Introduction

Sign language is the most frequently used tool when the transmission of audio is almost impossible or forbidden, or when the action of writing and typing is difficult, but the possibility of vision exists. Moreover, sign language is the most natural and expressive way for the hearing impaired.

Since sign language is gesticulated fluently and interactively like other spoken languages, a sign language recognizer must be able to recognize continuous sign vocabularies in real-time. The authors try to build such a system for the Taiwanese Sign Language. Two basic definitions used through out this paper are first given below.

Gestures are usually understood as hand and body movement which can pass information from one to another. Since we are interested in hand gesture and so the term "gesture" is always referred to the hand gesture in this paper.

Posture A posture is a specific configuration of hand flexion observed at some time instance.

Gesture A gesture is a sequence of postures connected by motions over a short time span.

Usually a gesture consists of one or more postures sequentially occurring on the time axis. In this paper, representing the semantic component in sign language, "gesture", "word", and "vocabulary" are all equivalent for convenience of description. Furthermore, a *sentence* is defined as a sequence of gestures in this paper.

2. Previous work

Fels' Glove Talk [3] focused on a gesture-to-speech interface. Moreover, a multilayer perceptron model was used in Beale and Edwards' posture recognizer [1] to classify sensed data into five postures in ASL. To help people with disabilities, Newby worked on the recognition of the letters and numbers of the ASL manual alphabet based upon statistical similarity [5].

A simplified method, using approximate spline, was proposed by Watson [1]. Gestures are represented by a sequence of critical points (local minima and maxima) of the motion of the hand and wrist[6][7]. This approach is more flexible in matching a gesture both spatially and temporally and thus reduces the computational requirement.

Starner and Pentlands' American Sign Language system [8][9][10] could recognize short sentences of American Sign Language (ASL) with 40 vocabularies, each was attached with its part of speech, which greatly reduced the computational complexity. The feature vector was fed to a hidden Markov model (HMM) for recognition of the signed words. This system gracefully integrated a useful concept in computational linguistics into gesture recognition. Furthermore, Nam's system [11] tried to recognize hand movement patterns. A HMM-based

method for recognizing the space-time hand movement pattern was proposed, and 10 kinds of movement primes could be recognized successfully.

Liang and Ouhyoung proposed a sign language recognition system [12] using hidden Markov model and integrated statistical approach used in computational linguistics. This system intended to recognize large set of vocabularies in a sign language by recognizing constructive postures and context information.

The system described in this paper is an extension of the one described above. The position, orientation, and motion model, in addition to the posture model, are implemented to enhance the performance of the system.

3. Taiwanese Sign Language

A vocabulary in Taiwanese Sign language can be defined as a gesture consisting of one or more postures, which has its own complete semantics. Similar to Stokoe's analysis of American Sign Language [13], four of the parameters are posture, position, orientation, and motion.

There are 51 fundamental postures in Taiwanese Sign Language (Appendix Figure A.1). Most gestures mainly contain only one posture, for example, *I*, *you*, *who*, etc., while gestures with multiple postures are also used, such as, *originally*, *father*, *mother*, *thank* and *good-bye*.

There are 22 typical positions that are often considered in Taiwanese Sign Language (Appendix Figure A.2). For example, vocabulary "I" is gesticulated as making posture number 1 (index finger) toward position number 7 (nose), while "eye" is the same except toward position number 6 (eye).

Gestures with different orientation usually indicate different objects in sign language. For instance, vocabulary "you two" and "we two" only differ on the orientation of the palm. Six different orientations of index finger are often used in TWL: up, down, right, left, ahead, and back.

Motion trajectory of a gesture also plays an important role to make the classification. For example, the parameters of vocabulary "thousand" and "ten thousand" are almost the same except in motion trajectories. Eight kinds of motions (Appendix figure A.3) need to be classified in the first three lessons of the TWL textbook.

Note that to integrate the above four parameters into recognition process, the end point problem must be first solved.

4. End-point problem

To determine end points in a sequence of gesture input, discontinuities are detected for segmentation. The discontinuity detection is done by time-varying parameter (TVP) detection; whenever the number of TVPs of the hand flexion begins to reduce to below a threshold, the motion of posturing is thought to be quasi-stationary, and its corresponding frame of data is taken to be recognized. Also, a filter is used to tolerate the jittering sensed data.

A gesture input stream can be thought as repeated patterns of the following states: transition and posture holding (Figure 1). On detecting the beginning of posture holding, the system extracts features, including position, orientation, and posture, and meanwhile, starts tracking motion trajectory. Until next transition happens, the motion trajectory tracked is analyzed. At this moment, all four parameters are available to perform a higher level gesture match, which is described in Section 6.

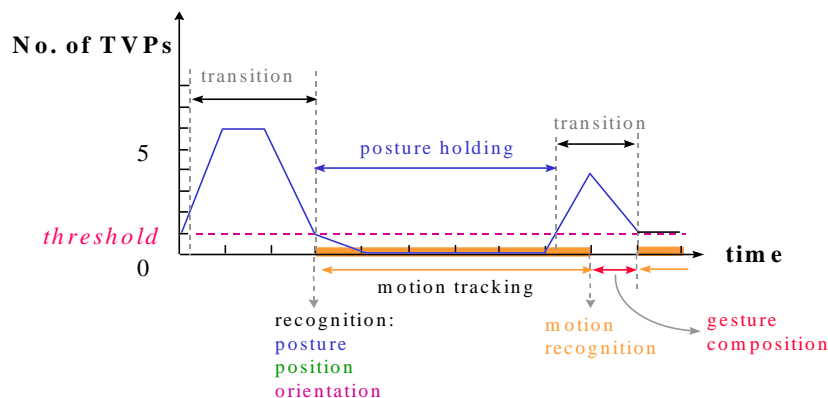


Figure 1. The number of TVPs of a gesture input stream varies along the time axis.

5. Recognition strategy

Consider an ending frame e , as illustrated in Figure 2. Each frame may be recognized as several posture

candidates in posture recognition phase and are composed into some gestures according the lexicon of Taiwanese Sign Language. Then, the typical dynamic programming is used to evaluate three conditions in Figure 2.

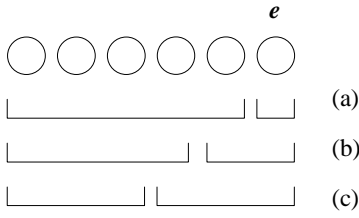


Figure 2. Possible gestures ending at frame e , with gesture of length 1(a), length 2 (b) and length 3(c).

To solve this dynamic programming problem, we first let $Solution(e)$ represent the best solution ending at frame e . We assume that g_1 is the gesture in Figure 2(a), g_2 in Figure 2(b), and g_3 in Figure 2(c); g_{1l} is the last gesture in $Solution(e-1)$, g_{12} is the last gesture in $Solution(e-2)$, and g_{13} is the last gesture in $Solution(e-3)$. The probability of the best solution ending at frame e can be written as

$$P(Solution(e)) = \max (P(Solution(e-1)) a_{g_{1l}g_1}P(g_1), P(Solution(e-2))a_{g_{12}g_2}P(g_2), P(Solution(e-3))a_{g_{13}g_3}P(g_3)). \quad (1)$$

where a_{ij} is the probability that gesture i and gesture j are adjacent and is called the *grammar model*. $P(g_1)$ is the probability of gesture g_1 in the specific sign language system and is called the *language model*.

6. System overview

Similar to posture model, the other three models are

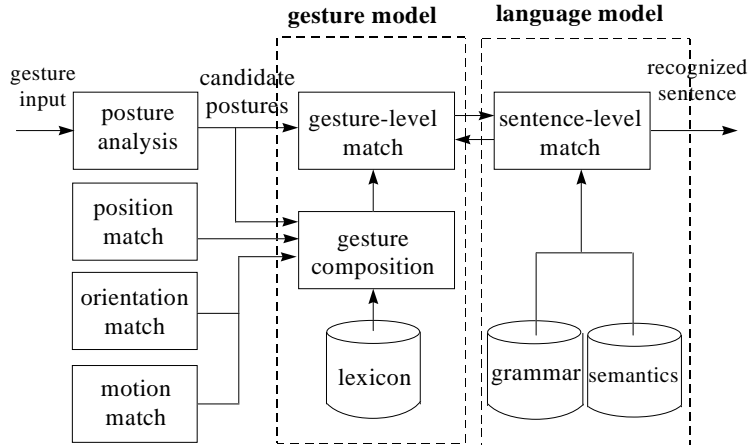


Figure 3. Block diagram of posture-based continuous gesture recognition.

6.1. Feature Extraction

The following features of a hand gesture are extracted as input of hidden Markov models for posture, orientation, and motion respectively.

- Flexion of 10 finger joints of one hand sensed by DataGlove is taken for posture recognition

triggered by discontinuity detection (by monitoring the number of TVPs). The proposed system architecture is shown in Figure 4. Posture analysis is illustrated in Figure 5 and is described later. After posture analysis, the results are decoded into several candidate postures. The gesture composition composes several possible gestures. Gesture-level match evaluates these gestures according to the probabilities of associated postures, positions, orientation, motion, and their corresponding probabilities in this language.

In Figure 3, the two arrows between gesture-level and sentence-level matches indicate the necessary backward and forward processes of dynamic programming. The relationship of several adjacent gestures is explained from storage grammar. The probability looked up from storage grammar is combined with the probability in gesture-level match, and the sentence-level match will generate a sentence with the highest probability and output it according to semantics. Thus, the evaluation of a certain vocabulary can be formulated as a weighted sum of the entropies (definition of entropy can be found in [16]) of the probabilities described above.

$$E(v) = w_{pr} * E_{pr}(v) + w_{ps} * E_{ps}(v) + w_o * E_o(v) + w_m * E_m(v) + w_u * E_u(v) + w_b * E_b(v_0, v) \quad (2)$$

where w_{pr} represents the weight of posture and $E_{pr}(v)$, the entropy of posture probability of vocabulary v . Similarly, w_{ps} , w_o , w_m , w_u , and w_b stand for weights of position, orientation, motion, uni-gram, and bi-gram respectively; v_0 is the vocabulary that precedes v .

- Azimuth, elevation, and roll of palm reported by a Polhemus 3D tracker are used for orientation recognition.
- A motion trajectory is normalized and divided into 10 connected vectors. The relative cosine of each pair of adjacent vectors, the number of turning points in the trajectory, and the relative orientation between start point

and end point of the motion path, are used for motion recognition.

7. Results

The system is implemented on an Pentium PC -133 PC. 250 vocabularies and 196 sentences were included for lexicon building and statistical learning respectively. The following four models all use left-right hidden Markov models, with the feature vectors listed in the previous subsection as input of HMMs.

Posture. Totally 613 posture samples are collected for 51 kinds of basic postures in training phase and other 281 samples are collected for test. The recognition rate is 95%, and the coverage rate (3 candidates) is 100%.

Position. Position model is somewhat complex to implement, however, we roughly classify positions into two parts: the one above the jaw and the one below. This reduces 64 pairs of indistinguishable gestures to only 3 pairs.

Orientation. 143 orientation samples are collected for 6 orientations in orientation training phase. The recognition rate is 90.1% and the coverage rate (3 candidates) is 100%, resulted from 71 test samples.

Motion. A total number of 279 motion trajectories are collected for motion training. 40 test motions are classified into one of the eight kinds of motions (Appendix Figure A.3). The recognition rate is 87.5% and the coverage rate is 100%.

Isolated gesture. Tests are divided into three parts: Lesson 1, including 71 vocabularies, Lesson 1 to Lesson 2, including 155 vocabularies, and Lesson 1 to Lesson 3, including 250 vocabularies.

	# of vocabularies 71	# of vocabularies 155	# of vocabularies 250
top 1 choice	84.5%	82.1%	70.5%
top 2 choices	94.4%	95.2%	88.4%
top 3 choices	97.2%	97.6%	89.5%

Table 1. Recognition rate and coverage rate (3 candidates included) of isolated gestures.

Sentence. Evaluation of sentence recognition rate

includes two parts, short sentence (Table 2) and long sentence (Table 3). 96 short sentences (2.41 words per sentence in average) in Lesson 1 are inside-tested, and the result is shown in the second column of Table 2. Furthermore, in Lesson 2, 108 short sentences (2.83 words per sentence in average) are inside-tested, and in Lesson 3, 99 short sentences (2.97 words per sentence in average) are inside-tested.

	# of vocabularies 71	# of vocabularies 155	# of vocabularies 250
top 1 choice	87.0%	72.5%	67.4%
top 2 choices	92.2%	83.3%	70.4%
top 3 choices	93.5%	86.2%	70.4%

Table 2. Recognition rate and coverage rate (3 candidates included) of short sentences.

Long sentence inside-tests are also done through 3 lessons (Table 3). 87 long sentences (5.24 words per sentence in average) in Lesson 1 are tested; 120 long sentences (4.98 words per sentence in average) in Lesson 2 and 138 long sentences (4.02 words per sentence in average) in Lesson 3 are also tested.

Summary of result. Table 4 shows average results through 3 lessons in TWL textbook in the above subsections.

Although 100% coverage rate can be achieved in posture, orientation, and motion models, the recognition rate of isolated gestures is 94.8%. This may result from the absence of the information from the other hand. About 24% of the vocabularies in the 3 lessons need the left hand to manipulate.

Keeping 3 candidate solutions for each time frame and tuning the weighting factors in Equation (14) by hand, the recognition rate for a short sentence (2.66 words per sentence in average) is 75.4%, and recognition rate for a long sentence (4.67 words per sentence in average) is 84.7%. The above result is inside tested from 303 test sentences for the short sentences and 345 sentences for the long sentences. Therefore, the weighted recognition rate is 80.4% and is 85.7% if top 3 choices are considered.

	# of vocabularies 71	# of vocabularies 155	# of vocabularies 250
top 1 choice	93.3%	84.4%	79.5%
top 2 choices	96.0%	87.4%	81.6%
top 3 choices	96.0%	88.9%	81.6%

Table 3. Recognition rate and coverage rate (3 candidates included) of long sentences.

As discussed in the previous section, only one 3D tracker on the palm makes the position model difficult to implement, since a signer may change the pose of body during gesticulation. Only if the information of a signer's body is available, either by multiple 3D trackers mounted on the user's body or by camera, can the position model be realized.

Furthermore, some gestures in TWL require both hands to manipulate in the same time and most of them use two different posture primes. This may be solved by using two DataGlove™s, and then applying the proposed model for recognition.

8. Future work

	Posture	Position	Orientation	motion	isolated gesture	short sentence	long sentence
top 1 choice	95%	*	90.1%	87.5%	78.4%	75.4%	84.7%
top 3 choices	100%	*	100%	100%	94.4%	83.4%	87.8%

Table 4. Recognition rate and coverage rate (3 candidates included) of posture, orientation, motion, and sentence recognition, where "*" indicates the incomplete implementation of position recognition module.

Acknowledgment: This project was partially supported by National Science Council under grant NSC-830425E002 140 and NSC-830408E002 006.

References

- [1] R. Watson. A Survey of Gesture Recognition Techniques. technical report TCD-CD-93-11, Department of Computer Science, Trinity College, Dublin 2, 1993.
- [2] R-H. Liang and M. Ouhyoung. A Real-time Continuous Alphabetic Sign Language to Speech Conversion VR System. *Computer Graphics Forum*, pp.C67-C77, Vol. 14, No. 3, UK, Aug 1995. (also in EUROGRAPHICS'95, Holland).
<http://www.cmlab.csie.ntu.edu.tw/~f1506028>.
- [3] S.S. Fels and G.E. Hinton. Building Adaptive Interfaces with Neural Networks: The Glove-talk Pilot Study. pp. 683-688, *Human-Computer Interaction- INTERACT'90*, IFIP, Elsevier Science Publishers B. V. (North-Holland), 1990.
- [4] K. Väänänen and K. Böhm. Gesture Driven Interaction as a Human Factor in Virtual Environments - An Approach with Neural Networks. pp. 93-106, *Virtual Reality System*, Academic Press, 1993.
- [5] G.B. Newby. Gesture Recognition Based upon Statistical Similarity. pp. 236-243, *Presence*, Vol. 3, No. 3, MIT Press, 1994.
- [6] R. Watson and P. O'Neill. A Flexible Gesture Interface. *Proc. of Graphics Interface '95*, Montreal, Canada, May 1995.
- [7] R. Watson and P. O'Neill. Gesture Recognition for Manipulation in Artificial Realities. *Proc. of the 6th International Conference on Human-Computer Interaction*, Pacifico Yokohama, Yokohama, Japan, July 1995.
- [8] T. Starner. *Visual Recognition of American Sign Language Using Hidden Markov Models*. Master's thesis MIT Media Lab, 1995.
<ftp://whitechapel.media.mit.edu/pub/tech-reports/TR-316.ps.Z>.
- [9] T. Starner and A. Pentland. Visual Recognition of American Sign Language Using Hidden Markov Models. Technical Report TR306, Media Lab, MIT, 1995.
<ftp://whitechapel.media.mit.edu/pub/tech-reports/TR-306.ps.Z>.
- [10] T. Starner and A. Pentland. Real-time American Sign Language Recognition from Video Using Hidden Markov Models. Technical Report TR375, Media Lab, MIT, 1996.
<ftp://whitechapel.media.mit.edu/pub/tech-reports/TR-375.ps.Z>.
- [11] Y. Nam and K-Y. Wohn. Recognition of Space-Time Hand-Gestures using Hidden Markov Model. pp. 51-58, *Proc. of the ACM Symposium on Virtual Reality Software and Technology*, Hong Kong, July 1996.
<http://vr.kaist.ac.kr/~nyh/nyh.html>.
- [12] R-H. Liang and M. Ouhyoung. A Sign Language Recognition System Using Hidden Markov Model and Context Sensitive Search. pp. 59-66, *Proc. of the ACM Symposium on Virtual Reality Software and Technology*, Hong Kong, July 1996.
<http://www.cmlab.csie.ntu.edu.tw/~f1506028>.
- [13] W.C. Stokoe. *Sign Language Structure*, Buffalo: University of Buffalo Press, 1960.
- [14] L.R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. pp. 267-296, *Proc. of the IEEE*, Vol. 77, No. 2, 1989.
- [15] L.R. Rabiner and B. Juang. *Fundamentals of Speech Recognition*, Prentice Hall International editions, U.S., 1993.
- [16] E. Charniak. *Statistical Language Learning*, MIT Press, Cambridge, Massachusetts, 1993.

Appendix

0. zero	1. one	2. two	3. three	4. four	5. five	6. six
7. seven	8. eight	9. nine	10. ten	11. twenty	12. thirty	13. forty
14. eighty	15. hundred	16. thousand	17. ten-thousand	18. female	19. hand	20. rectangle
21.	22. brother	23. people	24. together	25. keep	26. male	27. Lu
28. sister	29. tiger	30. fruit	31. nonsense	32. very	33. airplane	34. Chih
35. fist	36. borrow	37. gentle	38. subordinate	39. brown	40. boy scouts	41. vegetable
42. pen	43. similar	44. duck	45. money	46. dragon	47. worm	48. arm
49.	50. WC					

Figure A.1. Fifty-one fundamental postures in Taiwanese Sign Language, where No. 27 and 34 are translated from Chinese. No. 27 means a last name, and No. 34 is a quantity auxiliary.

	1. In front of the body		7. Nose		13. Neck		19. Under the arm
	2. Above the head		8. Mouth		14. Shoulder		20. Arm
	3. In front of the face		9. Jaw		15. Heart		21. The back of the hand
	4. Top of head		10. Temple		16. Breast		22. Wrist
	5. Brow		11. Ear		17. Waist		
	6. Eye		12. Cheek		18. Leg		

Figure A.2. Twenty-two basic gesture/body relative positions in Taiwanese Sign Language(TWL).

linear movement	circular movement	U-like movement	L-like movement
J-like movement	arm waving	wrist waving	wrist rotation

Figure A.3. Eight motion types in Taiwanese Sign Language(TWL).